DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 9 | Issue 2 Article 18

11-1-2010

Robust Estimators in Logistic Regression: A Comparative Simulation Study

Sanizah Ahmad
saniz924@salam.uitm.edu.my

Norazan Mohamed Ramli
Universiti Teknologi MARA (UiTM), Selangor Darul Ehsan, Malaysia, norazan@tmsk.uitm.edu.my

Habshah Midi
Universiti Putra Malaysia, Selangor Darul Ehsan, Malaysia

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 502-511

1538 – 9472/10/$95.00

Robust Estimators in Logistic Regression: A Comparative Simulation Study

Sanizah Ahmad Norazan Mohamed Ramli
Universiti Teknologi MARA (UiTM),
Selangor Darul Ehsan, Malaysia

Habshah Midi
Universiti Putra Malaysia,
Selangor Darul Ehsan, Malaysia

The maximum likelihood estimator (MLE) is commonly used to estimate the parameters of logistic regression models due to its efficiency under a parametric model. However, evidence has shown the MLE has an unduly effect on the parameter estimates in the presence of outliers. Robust methods are put forward to rectify this problem. This article examines the performance of the MLE and four existing robust estimators under different outlier patterns, which are investigated by real data sets and Monte Carlo simulation.

Key words: Logistic regression, robust estimates, downweighting, leverage points.

Introduction

Logistic regression models are widely used in the field of medical and behavioral sciences. These models are used to describe the effect of explanatory variables on a binary response variable. The logistic regression model assumes independent Bernoulli distributed response variables with the probability of a positive response modeled as

$$P\left(Y_i = 1 \middle| X = x_i\right) = F\left(x_i^T \beta\right)$$

where F is the logistic distribution function, $x_i \in \Re^p$ are vectors of explanatory variables and $\beta \in \Re^p$ is unknown. Such models are usually estimated by the maximum likelihood estimator (MLE) due to its efficiency under a parametric model. Unfortunately, the MLE is very sensitive to outlying observations.

Pregibon (1981) stated that the estimated parameters in logistic regression may be severely affected by outliers; hence, several robust alternatives which are much less affected by outliers are proposed in the literature (for example, Pregibon, 1981; Copas, 1988; Kunsch, et al., 1989; Carroll & Pederson, 1993; Bianco & Yohai, 1996; Croux & Haesbroeck, 2003). The goal of this article is to demonstrate a formal comparison between the MLE and several robust methods for logistic regression through a simulation study and real data examples.

Background

The logistic regression model assumes an independent Bernoulli response variable Y which takes values 1(for success) or 0 (for failure). Let $X = \left(1, x_1, ..., x_p\right)$ be a vector of independent explanatory variables. Given a binary variable Y and a $p \times 1$ vector X of covariates, the logistic regression model is of the form:

$$P\left(Y = 1 \middle| X = x_i\right) = F\left(x_i^T \beta\right) = \frac{exp\left(x_i^T \beta\right)}{1 + exp\left(x_i^T \beta\right)},$$

$$i = 1, ..., n \qquad (1)$$

Sanizah Ahmad is a lecturer on the Faculty of Computer and Mathematical Sciences. Email him at: saniz924@salam.uitm.edu.my. Habshah Midi is an Associate Professor in the Department of Mathematics. Email him at: habshah@math.upm.edu.my. Norazan Mohamed Ramli is a lecturer on the Faculty of Computer and Mathematical Sciences. Email him at: norazan@tmsk.uitm.edu.my.

where $\beta^T = (\beta_0, \beta_1, ..., \beta_p)$ is a vector of parameters and F is assumed to be a continuous and increasing distribution function. For estimating the β parameters, the maximum likelihood estimator (MLE) is classically used and is defined by an objective function

$$\hat{\beta}_{MLE} = \arg\max_{\beta} \sum_{i=1}^{n} l(Y_i, x_i; \beta) \qquad (2)$$

where the log-likelihood contributions are

$$l(Y_i, x_i; \beta) = Y_i \ln F(x_i^T \beta) + (1 - Y_i) \ln[1 - F(x_i^T \beta)] \qquad (3)$$

which gives an asymptotically efficient procedure for estimating β. Alternatively, the MLE may be obtained by minimizing the deviance,

$$\hat{\beta}_{MLE} = arg\ min_{\beta} \sum_{i=1}^{n} D_i(\beta) \qquad (4)$$

where

$$D_i(\beta) = \{-Y_i \ln[P(x_i^T \beta)] - (1 - Y_i) \ln[1 - P(x_i^T \beta)]\}.$$

Differentiating (2) with respect to β results in the likelihood score equation

$$\sum_{i=1}^{n} [Y_i - F(x_i^T \beta)] x_i = 0. \qquad (5)$$

These equations are solved iteratively by using either the Newton-Raphson or Fisher Scoring method. It is important to point out that the MLE in logistic regression does not exist when the data has no overlap. The estimator can only be estimated if the data has overlap where the two parts of data given by the values of the dependent variable, $\{X = x_i | Y_i = 1\}$ and $\{X = x_i | Y_i = 0\}$ are not separated in the space of explanatory variables (Albert & Anderson, 1984). The MLE is asymptotically normal and is an efficient estimator, nonetheless, it is extremely sensitive to outliers and hence is said to not be robust. For this reason, several robust alternatives of the MLE have been created to remedy this problem.

Outliers in Logistic Regression

It is important to distinguish between the different cases of outlying observations in logistic regression. In a binary logistic model, outliers can occur in the *Y*-space, the *X*-space or in both spaces. For binary data, all the *y*'s are 0 or 1, hence an error in the *y* direction can only occur as a transposition 0→1 or 1→0 (Copas, 1988). This type of outlier is also known as residual outlier or misclassification-type error. An observation which is extreme in the design space *X* is called a leverage outlier or leverage point: a leverage point can be considered good or bad.

A good leverage point occurs when $Y = 1$ with a large value of $P(Y = 1 | x_i)$ or when $Y = 0$ with small value of $P(Y = 1 | x_i)$, and vice versa for a bad leverage point. Victoria-Feser (2002) showed that the MLE can be influenced by extreme values in the design space, and the case of misclassification errors has been studied by Pregibon (1982) and Copas (1988). Croux, et al. (2002) found that the most dangerous outliers, termed bad leverage points, are misclassified observations which are at the same time outlying in the design space of *x* variables.

Robust Estimators in Logistic Regression

In general, two alternative approaches to making MLE more robust in logistic regression exist. The first is based on weighting the likelihood score function in (5), the so-called Mallows-class (Mallows, 1975; Hampel, et al., 1986, §6.3). Two types of estimators fall in this category: the Mallows-type and Schweppe-type estimators. The former were introduced by Kunsch, et al. (1989) where the weights depend on the response as well as the covariates. Mallows-type estimators were also suggested by Kunsch, et al. (1989) but were analyzed more deeply by Carroll and Pederson (1993). This type of estimator downweights in terms of the relative position in the design space (leverage) and often uses Mahalanobis distance.

A general robust estimate for the logistic model (1) is given by the solution in β of (Carroll and Pederson, 1993)

$$\sum_{i=1}^{n} w_i x_i \left\{ Y_i - F\left(x_i^T \beta\right) - c\left(x_i \beta\right) \right\} = 0, \quad (6)$$

with w_i being the weights which may depend on x_i, y_i, or both and $c(x_i, \beta)$ is a correction function defined to ensure consistency. If $w_i \equiv 1$ and $c(x_i, \beta) = 0$, then (6) gives the usual logistic regression estimate. If $w_i = w\left(x_i, x_i^T \beta\right)$ and $c(x_i, \beta) = 0$, then the weights depend only on the design, and the estimator is called Mallows class. The estimator thus represents a weighted maximum likelihood estimator. Stefanski (1985) suggested downweighting via robust Mahalanobis distance for the covariate vector, $\boldsymbol{x}$. If $w_i = w\left(x_i, x_i^T \beta, Y_i\right)$, then the estimator is in the Schweppe class (Kunsch, et al., 1989) where the weights depend on the response as well as the covariates. This estimator is also known as the conditionally unbiased bounded influence function (CUBIF) estimator.

The second robust approach is proposed by Pregibon (1982) who worked directly with the objective function in (4). He replaced the deviance function in (4) with a robust estimator defined by

$$\beta = \underset{\beta}{arg\,min} \sum_{i=1}^{n} \lambda \left[D_i \left(x_i^T \beta, y_i \right) \right], \quad (7)$$

where λ is a strictly increasing Huber's type function. This estimator was designed to give less weight to observations poorly accounted for by the model, however, this estimator did not downweight influential observations in the design space and was not consistent. Bianco and Yohai (1996) improved this method which was consistent and more robust than Pregibon's estimator by defining

$$\beta = \underset{\beta}{arg\,min} \sum_{i=1}^{n} \left\{ \rho \left[\begin{array}{l} D\left(x_i^T \beta, y_i\right) + G\left(F\left(x_i^T \beta\right)\right) \\ + G\left(1 - F\left(x_i^T \beta\right)\right) \end{array} \right] \right\}. \quad (8)$$

The ρ chosen by Bianco and Yohai (1996) is a bounded, differentiable and a nondecreasing function defined by

$$\rho(x) = \begin{cases} x - \left(x^2 / 2k\right) & \text{if } x \le k \\ k/2 & \text{otherwise} \end{cases} \quad (9)$$

where k is a positive number, $G(x) = \int_0^x \psi(-\ln u)\, du$ and $\psi(x) = \rho'(x)$ but stressed that other choices of ρ are possible. Croux and Haesbroeck (2003) extended the Bianco and Yohai estimator (BY) by including weights to the BY estimator to reduce the influence of outlying observations in the covariate space. This weighted BY (WBY) estimator, also called the Croux and Haesbroeck (CH) estimator, can be defined as:

$$\beta = \underset{\beta}{arg\,min} \sum_{i=1}^{n} w\left(x_i\right) \left\{ \rho \left[\begin{array}{l} D\left(x_i^T \beta, y_i\right) + G\left(F\left(x_i^T \beta\right)\right) \\ + G\left(1 - F\left(x_i^T \beta\right)\right) \end{array} \right] \right\} \quad (10)$$

where the weights $w(x_i)$, in order to be a decreasing function of robust Mahalanobis distances, are distances computed using the Minimum Covariance Determinant (MCD) estimator (see Rousseeuw & Leroy, 1987) taken as:

$$w\left(x_i\right) = \begin{cases} 1 & \text{if} \quad RD_i^2 \le \chi^2_{p,0.975} \\ 0 & \text{else} \end{cases}. \quad (11)$$

This WBY estimator remains consistent because the weighting is only used on the *x*-variables. Unfortunately, the above weighting procedure also reduces the weight of the good leverage points, which is not necessary and may lead to a loss of efficiency.

Methodology

Simulation Study

A simulation study was carried out to compare the robustness of the estimators discussed. These estimators are: the MLE and the four robust estimators for logistic regression, the Mallows-type estimator (MALLOWS) with weights depending on a robust Mahalanobis distance (Carroll & Pederson, 1993) and the conditionally unbiased bounded influence (CUBIF) estimator (Kunsch, et al., 1989), both of which are computed by standard available routines in the Robust package of S-Plus, the Bianco & Yohai (BY) estimator (1996) with choice of objective function and implementation (Croux & Haesbroeck, 2003) and the weighted Bianco-Yohai (WBY) estimator, both S-plus programs available at www.econ.kuleuven.be/public/NDBAE06/programs/.

Following the simulation study carried out by Croux and Haesbroeck (2003), a logistic regression model is generated with two independent normally distributed covariates. The error terms ε_i are drawn from a logistic distribution defined as:

$$y = I\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \geq 0\right).$$

The true parameter values are $\beta = (0, 2, 2)$ with sample size $n = 200$; a large sample size is chosen to avoid separation problems.

The simulation study is reported under a variety of situations. Initially, data without contamination, having two explanatory variables independently and normally distributed with zero mean and unit variance is considered. Next, to examine the robust properties of all, the data is contaminated in three different ways, similar to the idea proposed by Victoria-Feser (2002). First, proportions (a certain percentage) are taken of the responses y chosen randomly and changed from either 0 to 1 or 1 to 0; this constitutes the misclassification-type error. For each contaminated case, 1%, 3%, 5%, 7% and 10% of the original data set are contaminated. Second, the same proportions are taken to contaminate both covariates and replace them by the value of 2 for moderate leverage points. The same process is then repeated and replaces the value by 6 for extreme leverage points. Finally, the same proportions are considered and the generated data are contaminated with both types of outliers simultaneously which constitutes bad leverage points.

To further investigate leverage points, following the idea suggested by Bondell (2005), the proportions of the explanatory variables x_1 and x_2 were taken simultaneously and their values were replaced with $x = 1,...,7$ gradually from moderate to extreme covariates in the design space with $Y = 1$. The proportions of the observations with bad leverage points were then contaminated by replacing the explanatory variables with values $x = 1,...,7$ gradually with response variable $Y = 0$.

The five methods were then applied to these data under different situations already mentioned. In each simulation run included 1,000 replications. The performances of the five methods are evaluated based on the bias and the mean squared error (MSE). The bias for each parameter and the mean squared error are respectively defined as:

$$Bias = \left\| \frac{1}{1000} \sum_{i=1}^{1000} \hat{\beta}_i - \beta \right\|$$

and

$$MSE = \left(\frac{1}{1000} \sum_{i=1}^{1000} \left\| \hat{\beta}_i - \beta \right\|^2 \right)$$

where $\|\cdot\|$ indicates the Euclidean norm (Croux & Haesbroeck, 2003).

Results

The bias and the MSE of the five estimates are shown in Tables 1-5. A good estimator is one that has bias and MSE which are relatively small or close to zero. It can be observed that, in clean data with zero percentage of outliers, the biases and MSEs of all five estimators are fairly close to each other.

Table 1 shows data with misclassified errors. The bias and MSE of the MLE estimates were immediately affected by 1% misclassified-type error. The results suggest that the MLE

becomes biased with 1% contamination, CUBIF with 3% contamination and BY with 7% contamination. The MALLOWS and WBY exhibit good robust estimators with the latter being the best method.

It can be observed from Table 2 that there is not much difference between the classical and the robust methods when contaminating data with extreme leverage points (replacing x by 6 and $Y = 1$). Similar results were obtained for moderate leverage points (replacing x by 2 and $Y = 1$); these results are not shown due to space limitations.

It is interesting to observe the results of Table 3 in the situation where 5% of the data was contaminated with leverage points by gradually increasing the distance of x. Similar conclusions to those from Table 2 can be made where the biases and MSEs for all methods are relatively small. Hence, it can be concluded that leverage points do not have much effect on the data because this type of contamination is considered as good leverage points.

Table 1: Bias and MSE of All Methods for Data with Various Percentages of Misclassified Errors

% of	MLE		MALLOWS		CUBIF		BY		WBY	
misc error	bias	MSE	bias	MSE	bias	MSE	bias	MSE	bias	MSE
0	0.0909	0.2781	0.0871	0.2774	0.0898	0.2782	0.1074	0.3089	0.1088	0.3204
1	1.4570	2.1918	0.1400	0.3073	0.5148	0.4482	0.1344	0.3765	0.1745	0.3842
3	2.4648	6.1013	0.3484	0.2098	1.1933	1.4598	0.2542	0.1716	0.0565	0.1120
5	2.7288	7.4773	0.4309	0.6467	1.6603	2.8031	0.7688	1.3257	0.0703	0.3217
7	2.8247	8.0053	0.4354	0.5614	2.0318	4.1658	2.8258	8.0112	0.3752	0.5560
10	2.8838	8.5320	0.7716	0.8849	2.4287	5.9337	2.8771	8.3148	0.0515	0.3961

Table 2: Bias and MSE of All Methods for Data with Various Percentages of Extreme Leverage Points

% of	MLE		MALLOWS		CUBIF		BY		WBY	
lev pt	bias	MSE	bias	MSE	bias	MSE	bias	MSE	bias	MSE
0	0.0840	0.2616	0.0817	0.2588	0.0833	0.2605	0.0845	0.2812	0.0875	0.2908
1	0.8072	0.8122	0.8115	0.8188	0.8122	0.8219	0.8138	0.8356	0.8198	0.8513
3	0.8035	0.8096	0.8083	0.8183	0.8060	0.8143	0.8121	0.8416	0.8118	0.8506
5	0.7910	0.7903	0.7954	0.7979	0.7911	0.7922	0.8019	0.7867	0.7867	0.8101
7	0.8089	0.8392	0.8124	0.8452	0.8111	0.8442	0.8150	0.8601	0.8109	0.8632
10	0.8162	0.8421	0.8216	0.8519	0.8186	0.8458	0.8454	0.9045	0.8463	0.9096

Table 3: Bias and MSE of All Methods for Data with 5% Leverage Points for Various Distances

	MLE		MALLOWS		CUBIF		BY		WBY	
distance	bias	MSE	bias	MSE	bias	MSE	bias	MSE	bias	MSE
clean	0.0909	0.2781	0.0871	0.2774	0.0898	0.2782	0.1074	0.3089	0.1088	0.3204
x=1	0.4882	0.4206	0.4909	0.423	0.4894	0.4233	0.5019	0.4601	0.4990	0.4758
x=2	0.5037	0.4219	0.5060	0.4255	0.5050	0.4243	0.5027	0.4366	0.4963	0.4434
x=3	0.5479	0.4861	0.5527	0.4923	0.5528	0.4929	0.5319	0.4824	0.5342	0.4887
x=4	0.5175	0.4495	0.5220	0.4549	0.5195	0.4518	0.5154	0.4639	0.5211	0.4732
x=5	0.4987	0.4156	0.5022	0.4183	0.5010	0.4178	0.4853	0.4191	0.4848	0.4275
x=6	0.4873	0.4059	0.4894	0.4087	0.4885	0.4093	0.4776	0.4119	0.4768	0.4160
x=7	0.4612	0.3972	0.4635	0.4004	0.4633	0.4009	0.4279	0.3925	0.4249	0.4103

The presence of moderate and extreme bad leverage points changes the picture dramatically. It can be observed from Tables 4 and 5 that for both cases, the CUBIF estimator can only withstand up to 3% contamination. The BY estimator can tolerate up to 3% contamination when $x = 2$, and 5% contamination when $x = 6$. The WBY estimator is better than the MALLOWS for the moderate bad leverage points. In this situation, the WBY and the MALLOWS can only withstand up to 3% and 1% contamination, respectively. Nevertheless, with data having extreme bad leverage points, the performances of the WBY and MALLOWS are equally good: both estimators are able to withstand up to 10% contamination.

Finally the results shown in Table 6 are discussed in the context of the situation where the data has 5% bad leverage points and is at various distances of the explanatory variables. By gradually increasing the distance of x and when $Y = 0$, the MLE is biased for all x; the bias worsens as x increases for MLE, but bias is consistent with the CUBIF estimators. By contrast, the bias of the MALLOWS estimator is small for $x = 6$ and $x = 7$. The BY estimator performs best when the bad leverage points are located at $x = 5$ and $x = 6$. Conversely, the biases and MSEs of the WBY estimates are consistently the smallest among other estimators for $x = 2,\ldots,7$. The results shown in Table 6 reveal that the WBY performs much better compared to the other estimators.

Numerical Examples

Two real data sets are considered to illustrate the behavior of the various robust estimates discussed. Results of the estimated coefficients, as well as their standard errors, are presented for the original and the modified data. The modified data refer to the original data with deleted outlier observation(s). A good estimator is one that has parameter estimates reasonably close to the MLE estimates of the modified data (clean data). Kordzakhia, et al. (2001) suggested another criterion for evaluating various estimators. They proposed comparing the various estimates using a goodness-of-fit discrepancy, the Chi-square statistic based on the arcsin transformation χ^2_{arc} defined as

Table 4: Bias and MSE of All Methods for Data with Moderate Bad Leverage Points (Replacing *x* by 2 and Y=0)

% of	MLE		MALLOWS		CUBIF		BY		WBY	
bad lev pt	bias	MSE	bias	MSE	bias	MSE	bias	MSE	bias	MSE
0	0.0909	0.2781	0.0871	0.2774	0.0898	0.2782	0.1074	0.3089	0.1088	0.3204
1	0.6339	0.5457	0.4976	0.3976	0.4249	0.3403	0.1222	0.2839	0.0072	0.3100
3	1.4107	2.0720	1.2084	1.5427	1.0922	1.2793	0.5695	0.5144	0.1954	0.3150
5	1.8501	2.0720	1.6461	2.7746	1.5235	2.3883	1.0211	1.1926	0.3895	0.4337
7	2.1888	4.8457	2.0127	4.1041	1.9247	3.7592	1.6166	2.7169	0.6992	0.7607
10	2.3917	5.7686	2.2550	5.1330	2.2226	4.9893	2.1789	4.8230	1.0665	1.3894

Table 5: Bias and MSE of All Methods for Data with Extreme Bad Leverage Points (Replacing *x* by 6 and Y=0)

% of	MLE		MALLOWS		CUBIF		BY		WBY	
bad lev pt	bias	MSE	bias	MSE	bias	MSE	bias	MSE	bias	MSE
0	0.0909	0.2781	0.0871	0.2774	0.0898	0.2782	0.1074	0.3089	0.1088	0.3204
1	1.4570	2.1918	0.1400	0.3073	0.5148	0.4482	0.1344	0.3765	0.1745	0.3842
3	2.4648	6.1013	0.3484	0.2098	1.1933	1.4598	0.2542	0.1716	0.0565	0.1120
5	2.7288	7.4773	0.4309	0.6467	1.6603	2.8031	0.7688	1.3257	0.0703	0.3217
7	2.8247	8.0053	0.4354	0.5614	2.0318	4.1658	2.8258	8.0112	0.3752	0.5560
10	2.8838	8.5320	0.7716	0.8849	2.4287	5.9337	2.8771	8.3148	0.0515	0.3961

Table 6: Bias and MSE of All Methods for Data with Bad Leverage Points at 5% Contamination for Various Distances

	MLE		MALLOWS		CUBIF		BY		WBY	
distance	bias	MSE	bias	MSE	bias	MSE	bias	MSE	bias	MSE
clean	0.0909	0.2781	0.0871	0.2774	0.0898	0.2782	0.1074	0.3089	0.1088	0.3204
x=1	1.2243	1.5730	1.2404	1.6134	1.2344	1.5975	1.0517	1.2093	1.0817	1.2803
x=2	1.8718	3.5497	1.7615	3.1523	1.6518	2.7846	1.1034	1.3547	0.4069	0.7711
x=3	2.2447	5.0795	1.8442	3.4507	1.6346	2.7267	0.9045	1.0110	0.1705	0.3836
x=4	2.4888	6.2345	1.6528	2.8113	1.6403	2.746	0.7273	0.7795	0.1515	0.3691
x=5	2.6367	6.9921	1.1466	1.4881	1.6387	2.7385	0.5689	0.6169	0.1243	0.3584
x=6	2.7193	7.4377	0.4851	0.5108	1.6465	2.7669	0.5183	0.8591	0.1290	0.3492
x=7	2.7635	7.6605	0.2009	0.1815	1.6542	2.7695	1.2914	3.2433	0.1693	0.2076

$$\chi^2_{arc} = \sum_{i=1}^{n} 4\left[arcsin\sqrt{y_i} - arcsin\sqrt{\pi_i} \right]^2,$$

where $\sqrt{\pi_i}$ represents the fitted probabilities for $i = 1, 2, ..., n$. The lower χ^2_{arc}, the better the goodness-of-fit.

Example: Leukemia Data

The Leukemia Data (Cook & Weisberg, 1982) was analyzed by Carroll and Pederson (1993), among others. The data set consists of measurements on 33 leukemia patients. The response variable is 1 if the patient survived more than 52 weeks and 0 otherwise. Two covariates are present in the model: white blood cell count (WBC) and AG status, which is the presence or absence of certain morphologic characteristic in the white cells. Cook and Weisberg (1982) considered these data to illustrate the identification of influential observation and they detected one observation (#15), corresponding to a patient with WBC = 100,000 who survived for a long period of time to be influential when the MLE was used. The plot in Figure 1 suggests that the observation looks like a bad leverage point.

Table 7 exemplifies the estimated parameters and estimated standard errors for the various procedures including MLE32. The MLE32 refers to the MLE estimates for the clean data after deleting observation (#15). A good estimator is one that has parameter estimates fairly close to the MLE32. It can be observed from Table 7 that the MALLOWS and WBY estimates are reasonably close to the MLE32 estimates. However, the Mallows Chi-square statistic is larger than the WBY, hence, the WBY is the best estimator for Leukemia Data because it gives the smallest χ^2_{arc} value and their estimates are closer to the MLE32. WBY is followed by the MALLOWS, BY and CUBIF estimators.

Example: Vaso-Constriction Data

The Vaso-constriction data is a well-known dataset referred to as skin data. It was introduced by Finney (1947) and was studied by Pregibon (1982) to illustrate the impact of potential influential observations in logistic regression. The binary outcomes (presence or absence of vaso-constriction of the skin of the digits after air inspiration) are explained by two explanatory variables: x_1 the volume of air inspired, and x_2 the inspiration rate (both in logarithms). The literature, which extensively uses this dataset, often reports observations (#4) and (#18) as outliers. As shown in Figure 2, a plot of the data based on the maximum likelihood fit shows that the two observations (#4 and #18) look more like misclassified errors rather than outlying observations.

Table 8 presents the estimated parameters, estimated standard errors and goodness-of-fit measures for the various procedures including MLE37 after removing the two influential observations. Several interesting points appear from Table 8. It is notable that the

Figure 1: Scatter Plot of Leukemia Data

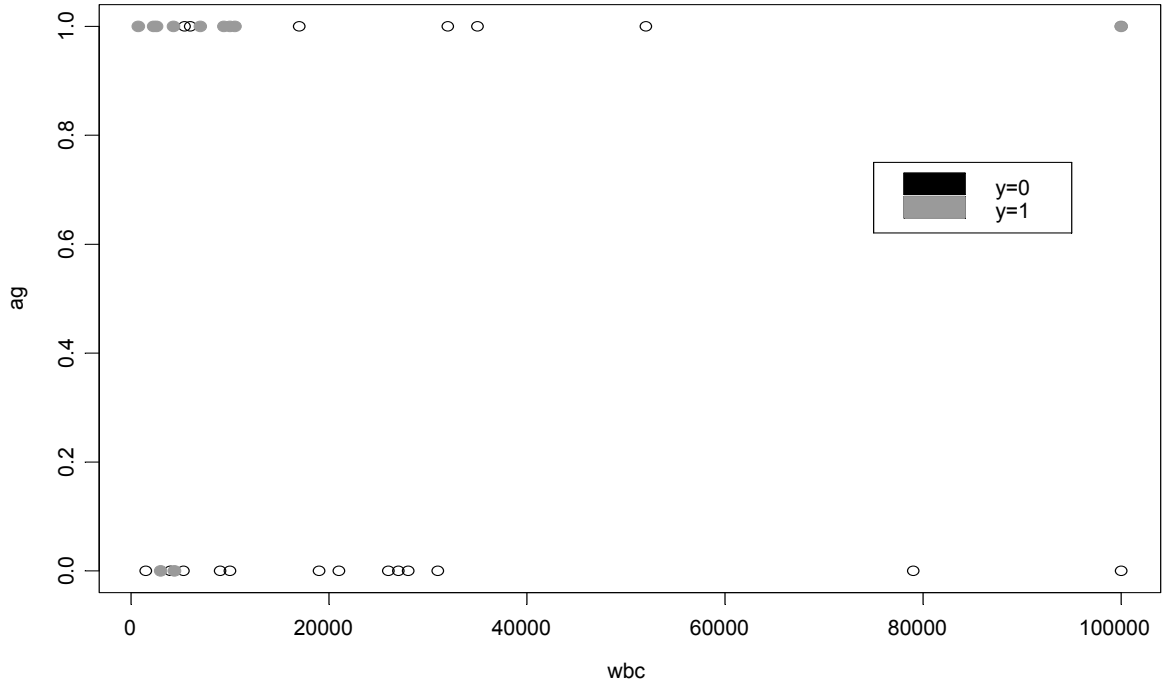

Table 7: Leukemia Data: Estimated Parameters, Standard Errors and Goodness-of-Fit Measures

Estimation	Intercept		WBC		AG		
Method	Est.	s.e.	Est.	s.e.	Est.	s.e.	χ^2_{arc}
MLE	-1.3073	0.2931	-0.3177	1.454	2.2611	2.2003	52.16
MLE32	0.2119	7.0996	-2.3545	6.9497	2.5581	4.9143	32.52
MALLOWS	0.1710	6.7568	-2.2535	6.7818	2.524	4.6589	42.46
CUBIF	-0.6763	1.7135	-0.9110	3.4500	2.2495	1.1712	46.73
BY	0.1595	5.0511	-1.7740	5.7623	1.9276	3.3011	44.05
WBY	0.1891	6.8884	-2.1927	6.7853	2.4003	4.6923	39.47

CUBIF and MALLOWS yield results reasonably close to the MLE. The results also show that the BY and WBY estimates have been strongly affected when the two influential observations are removed from the dataset. It may be observed that the parameter estimates and the standard errors of both estimates become large because, without the two observations, the remaining data set is in a situation of quasi-complete separation (Albert & Anderson, 1984), with little overlap between observations $y_i = 0$ and $y_i = 1$. Thus, the model is nearly undetermined. For this reason, the BY and WBY downweight these observations and have large increases of coefficients and standard errors. The parameter estimates and the standard errors of both estimators are considerably close to the MLE37 estimates. However, the BY has the smallest χ^2_{arc} value, therefore, the BY estimator gives the best result for this data set.

Figure 2: Scatter Plot of Vaso Data

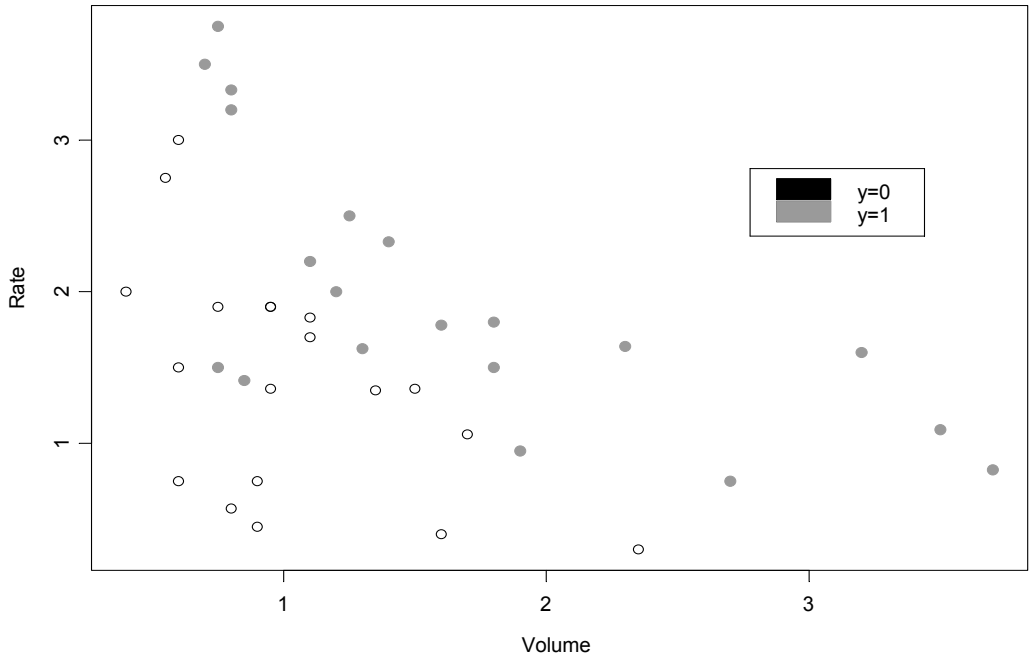

Table 8: Vaso Data: Estimated Parameters, Standard Errors and Goodness-of-Fit Measures

Estimation	Intercept		Log(Volume)		Log(Rate)		χ^2_{arc}
Method	Est.	s.e.	Est.	s.e.	Est.	s.e.	
MLE	-2.9239	1.2877	5.2205	1.8579	4.6312	1.7889	48.39
MLE37	-24.5812	14.0211	39.5498	23.2463	31.9352	17.7595	12.34
MALLOWS	-2.9207	1.2908	5.1673	1.8470	4.5967	1.7886	48.41
CUBIF	-2.8776	1.2707	5.1661	1.8364	4.5646	1.7644	48.47
BY	-6.8667	10.0507	10.7523	15.3086	9.381	12.7798	40.87
WBY	-6.8465	10.0672	10.7504	15.3346	9.3785	12.8014	40.91

Conclusion

The goal of this study was to compare the performance of the MLE and four robust estimators for the logistic model under both clean and contaminated data sets. The findings signify that the MLE can be biased in the presence of misclassified error and bad leverage points, whereas some robust estimators are better than others depending on the type of contamination. When the contamination data are leverage points, the simulation results indicate that all parameter estimates are not dramatically affected, because they have consistently small bias. Overall, the WBY estimator is preferred because it is more robust than other estimators tested in this study for any type of contamination in the data. This estimator is followed by the BY, MALLOWS and CUBIF. However, further investigation is needed to compare these robust estimators through an extensive simulation study involving different parameter values, sample sizes and parameter size. Further studies are also needed to investigate more suitable robust methods to cater outlying observations in logistic regression. Most robust methods unfortunately rely on simple downweighting of

distant observations in the design space regardless of whether or not they are misspecified, whether they are good or bad leverage points and what influence they have on the model.

References

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, *71*, 1-10.

Bianco, A. M., & Yohai, V. J. (1996). Robust estimation in the logistic regression model. In *Robust statistics, Data analysis and computer intensive methods,* H. Reider, Ed., 17-34. New York: Springer Verlag.

Bondell, H. D. (2005). Minimum distance estimation for the logistic regression model. *Biometrika*, *92*, 724-731.

Carroll, R. J., & Pederson, S. (1993). On robust estimation in the logistic regression model. *Journal of the Royal Statistical Society*, *Series B*, *55*, 693-706.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression.* London: Chapman and Hall.

Copas, J. B. (1988). Binary regression model for contaminated data (with discussion). *Journal of the Royal Statistical Society*, *Series B*, *50*, 225-265.

Croux, C., Flandre, C., & Haesbroeck, G. (2002). The breakdown behavior of the maximum likelihood estimator in the logistic regression model. *Statistics & Probability Letters*, *60*, 377-386.

Croux, C., & Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics & Data Analysis Journal*, *44*, 273-295.

Finney, D. J., (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika*, *34*, 320-334.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust Statistics: The approach based on influence functions*. New York, NY: John Wiley.

Kordzakhia, N., Mishra, G. D., & Reiersolmoen, L. (2001). Robust estimation in the logistic regression model. *Journal of Statistical Planning and Inference*, *98*, 211-223.

Kunsch, H. R., Stefanski, L. A., & Carroll, R. J. (1989). Conditionally unbiased bounded influence estimation in general regression models, with applications to generalized linear models. *Journal of American Statistical Association*, *84*, 460-466.

Mallows, C. L. (1975). *On some topics in robustness*. Murray Hill, NJ: Bell Telephone Laboratories.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, *9*, 705-724.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrika*, *73*, 413-425.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection.* New York: Wiley.

Stefanski, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika*, *72*, 583-592.

Victoria-Feser, M.-P. (2002). Robust inference with binary data. *Psychometrica*, *67*, 21-32.